Efficient Image Sampling in Diffusion Models via Rough Set Selection and Mamba State Spaces

Jianxing Yu, Zhiyong She*, Yijin Shi
Information Network Security College, Xinjiang University of Political Science and Law, Tumxuk, Xinjiang 844000, China
E-mail: szysoft378@outlook.com
*Corresponding author

Keywords: rough set, DDPM, Mamba, generative mode

Received: September 24, 2025

The Denoising Diffusion Probabilistic Model (DDPM) faces significant challenges, including low sampling efficiency, high computational complexity, and substantial resource demands when processing long sequences. To address these issues, this study introduces an innovative framework that integrates Rough Set theory with a Mamba-based state-space model (SSM).In our approach, each timestep in the diffusion process is treated as an object within a Rough Set domain, where temporal intervals define equivalence relations. These relations yield equivalence classes, and for any subset of the domain, we compute its roughness based on these classes. The optimal sampling sub-sequence is then selected by identifying the subset with minimal roughness, ensuring a more stable and representative sampling path compared to random strategies. During the model's training, the data object at each timestep initializes the input and state matrices of the SSM. The state transition from the previous timestep and the current input are computed dynamically based on the temporal intervals. This process allows the SSM to perform selective state updates, significantly enhancing the training efficiency of the proposed Mamba-DDPM. We evaluated our method against DDPM, DDIM, VAE-DDPM, and ViT-DDPM on the ImageNet and FFHQ datasets. Experimental results demonstrate the superiority of our approach across multiple metrics, including FID, SSIM, PSNR, and image generation time at various resolutions. Specifically, compared to ViT-DDPM, our method achieved relative improvements of 0.30%~15.39% in FID, 2.44%~21.13% in SSIM, and 0.35%~3.34% in PSNR for 128x128 image generation. For 512x512 generation, the gains were 2.12%~13.06% in FID, 1.27%~14.29% in SSIM, and 0.62%~2.14% in PSNR. We conclude that the proposed method effectively mitigates the inherent limitations of DDPM and outperforms other leading diffusion models.

Povzetek: Razvita Metoda z uporabo grobih množic izboljša učinkovitost in kakovost generiranja slik ter odpravlja glavne omejitve klasičnih difuzijskih modelov.

1 Introduction

At present, the two mainstream technologies in the field of generative models are antagonistic neural networks (GAN) [1] and diffusion models. Due to the inherent difficulty of convergence in training the generator and discriminator, the dominant position of GAN is gradually replaced by the diffusion model. Among them, the denoising diffusion probability model (DDPM) [2] proposed by Ho et al. and others has become a benchmark scheme in the diffusion model family. Through the progressive denoising mechanism, the model can not only capture more abundant generated details, but also generate high-quality images with high diversity. In the non-conditional image generation task, DDPM has shown the performance of surpassing GAN. For example, the research of Prafulla Dhariwal team in the field of image vision shows that the diffusion model has surpassed GANs in the image synthesis task [3].

Although DDPM excels in generation quality, its Markov chain-based long sequence sampling procedure results in lower efficiency than GAN. To address this limitation, a large number of optimization studies have been carried out in academia, aiming at improving the sampling speed while maintaining the generation quality. The non-Markov chain denoising diffusion probability model (DDIM) proposed by Nichol et al. [4] improves the efficiency through the subsequence sampling strategy, but the mechanism of randomly selecting the subsequence leads to the quality fluctuation of the generated results. In order to solve this problem, She et al. introduced rough set theory and proposed an improved denoising diffusion probability model [5], which achieved more stable generation quality while ensuring sampling efficiency.

Attention mechanism has been introduced to enhance the global modeling ability of U-net network to overcome the limitations of its local receptive field. Representative works include the U-net framework for medical image segmentation developed by Guu et al, which integrates spatial channel dual attention mechanism and transformer structure [6], and the scalable transformer diffusion model proposed by William and Xie [7]. These improved strategies enhance the ability to extract the key features of the image through the attention mechanism or

Transformer architecture. However, when Transformer is applied in the diffusion probability model (DDPM), the computational complexity of the model increases to the quadratic order, which significantly reduces the reasoning speed. Nevertheless, such technical improvements have promoted the in-depth application of DDPM in many frontier fields, including medical image analysis (object detection [8-10], lesion segmentation [11]), cross-media generation (image and video synthesis [12]) and multi-modal fusion technology [13].

Although the existing research has optimized the diffusion probability model (DDPM), its core defects have not been fundamentally solved [14]. It is noteworthy that the Mamba model developed by Gu et al. [15] achieves efficient sequence modeling with linear time complexity through selective state space mechanism, which shows significant advantages over traditional RNN and Transformer. But also can effectively deal with that memory problem of a long sequence and support parallel operation. Inspired by the technical path of Visual Transformer (ViT), the Visual State Space Model (VMamba) [16-20] proposed by Zhu et al. [16], Liu et al. [17] successfully extends the framework to the field of computer vision. In this study, a hybrid Mamba-DDPM architecture based on rough set theory is innovatively constructed: firstly, rough set theory [21-23] is used to intelligently select feature subsequences, and then the complexity and resource consumption in the subsequence sampling stage are significantly reduced by using the linear computing characteristics of Mamba. This interdisciplinary method, which combines rough set feature selection for stability, Mamba modeling for linear complexity, and DDPM generation for quality, provides a technical solution that addresses DDPM's inefficiencies in applications such as time series analysis and visual perception.

This approach holds potential application value in fields such as medical imaging, remote sensing, and real-time video generation. In medical imaging, for instance, in low-dose CT image denoising tasks, the method can preserve image details under noisy conditions (e.g., low signal-to-noise ratio), thereby improving diagnostic accuracy. Experimental results on public datasets such as LIDC-IDRI demonstrate that, under constrained conditions (e.g., limited computational resources), the proposed method achieves an improvement of approximately 15% in PSNR compared to traditional Denoising Diffusion Probabilistic Models (DDPMs). In remote sensing, the method has been applied to high-resolution satellite image restoration, such as land cover classification tasks. Under uncertain weather conditions (e.g., cloud occlusion), it significantly enhances image sharpness through rough set-optimized subsequences. Tests conducted on the UC-Merced dataset under simulated dynamic environments validate the robustness of the method in the presence of uncertainties. For real-time video generation, the method supports video super-resolution tasks on mobile devices. In computationally constrained environments (e.g., edge devices), it leverages the linear complexity of Mamba to achieve efficient processing, reducing the generation time by 30% compared to ViT-DDPM in practical tests. Future work will focus on further exploring the practical deployment of these application scenarios and evaluating the adaptability of the method in complex environments.

This method demonstrates significant application potential in domains such as medical imaging, remote sensing, and real-time video generation. In medical imaging, its efficient sampling capability can support real-time reconstruction and denoising of dynamic medical sequences, such as ultrasound imaging. In remote sensing image processing, the feature selection mechanism driven by rough set theory can effectively preserve critical spectral features for land cover classification. For real-time video generation scenarios, the linear complexity of Mamba models can meet the low-latency requirements of high-frame-rate synthesis. Furthermore, subsequent work in this paper will focus on enhancing the model's robustness to noisy or incomplete inputs. Inspired by methodologies from adaptive control and nonlinear control for handling system uncertainties, we will explore dynamic adjustment of state-space parameters—such as incorporating time-varying characteristics into the system matrix A—to improve model stability under partial observations or noisy conditions. This research direction is expected to significantly boost the model's applicability in practical scenarios, including clinical diagnosis with low-quality medical images and remote sensing image restoration under cloud occlusion.

The integration of Rough Set theory, Mamba, and DDPM is motivated by a systematic approach to address the core limitations of diffusion models. DDPM suffers from low sampling efficiency due to its long Markov chain-based denoising process. While DDIM improves speed through subsequence sampling, its random selection strategy leads to unstable generation quality. This creates a critical need for an intelligent subsequence selection mechanism that ensures both efficiency and representativeness.

Here, Rough Set theory provides a mathematical foundation for stable subsequence selection. By treating each timestep as an object and defining equivalence relations based on temporal intervals, Rough Set theory calculates the roughness of any candidate subsequence. Minimizing this roughness identifies an optimal, representative sampling path, effectively replacing DDIM's random strategy and ensuring a more stable and efficient denoising trajectory.

However, selecting an optimal subsequence is only one part of the solution. The denoising network itself, often a U-Net or Transformer, contributes significantly to computational complexity. The Mamba architecture, with its selective state space models (SSMs), is introduced to overcome this. Mamba offers linear computational complexity and efficient long-range dependency modeling, directly addressing the high resource consumption and limited receptive field of traditional denoisers.

Thus, the combination logic is clear and sequential: Rough Set theory first intelligently reduces the number of necessary sampling steps by selecting a minimal-roughness subsequence. Then, the Mamba model serves

as the high-efficiency denoising network operating on this shortened path, leveraging its linear-time state-space mechanisms. Finally, DDPM provides the proven probabilistic framework for high-quality image generation. This synergy creates a cohesive pipeline where Rough Sets ensure sampling stability, Mamba guarantees computational efficiency, and DDPM maintains generation fidelity, collectively mitigating the inherent bottlenecks of the original model.

A clear distinction is established between technical concepts and application scenarios. On the technical side, the discussion primarily centers on the Denoising Diffusion Probabilistic Model (DDPM) and its derivatives, including the Markov chain-based DDPM, the non-Markovian DDIM, U-Net and Transformer architectures incorporating attention mechanisms, as well as the selective state-space mechanism introduced by the Mamba model. These methods have progressively optimized the sampling efficiency and feature extraction capabilities of diffusion processes at the theoretical level.

On the application side, the focus extends across multiple frontier domains, such as object detection and lesion segmentation in medical imaging, cross-modal image and video synthesis, multimodal fusion analysis, and remote sensing image processing. By explicitly separating the discussion of core model mechanisms from that of practical application scenarios, this work not only highlights the theoretical advances of generative models but also demonstrates their broad adaptability to diverse visual tasks.

Within this research context, the present study aims to address two critical questions:

1) Rough Set-Based Sub-Sequence Selection Mechanism: By introducing rough set theory, a sub-sequence selection strategy is developed to replace traditional random sampling, thereby enhancing sampling stability.

2) Efficiency of the Mamba Architecture: Can the Mamba architecture significantly reduce the sampling complexity and resource consumption of DDPM without compromising image generation quality? By leveraging the linear computational complexity and efficient sequence modeling capabilities of the Mamba model, the denoising network of DDPM is reconstructed to ensure high-fidelity image generation while substantially improving sampling efficiency.

To this end, we propose an innovative hybrid Mamba-DDPM framework based on rough set theory. Specifically, rough set theory [21–23] is first employed to select informative feature sub-sequences, and subsequently, the linear computational characteristics of Mamba are utilized to markedly reduce the complexity and resource requirements during the sub-sequence sampling stage. This integration of rough set-based feature selection provides a reliable and efficient technical solution for diffusion models in applications such as temporal analysis, visual perception, multi-modal fusion, and remote sensing image processing.

How Mamba-DDPM Effectively Addresses These Limitations. Mamba-DDPM addresses the bottlenecks of traditional DDPM, such as low sampling efficiency, high computational complexity, and high resource consumption, by integrating rough set theory for intelligent subsequence selection, state-space models for linear-complexity sequencing, and diffusion models for denoising. Its linear-complexity architecture, combined with hardware-level optimizations, provides an efficient and high-quality solution for high-resolution image generation, paving the way for practical applications of diffusion models.

Limitations of Current State-of-the-Art Methods:

Low Sampling Efficiency: The iterative denoising mechanism of DDPM, based on a Markov chain, requires long-sequence sampling (typically $T \geq 1000$), resulting in inference speeds significantly lower than those of generative adversarial networks (GAN). Although DDIM improves efficiency through sub-sequence sampling, its random selection strategy often leads to unstable generation quality.

High Computational Complexity: Improved models incorporating attention mechanisms or Transformer architectures, such as ViT-DDPM, enhance feature extraction capabilities but incur quadratic computational complexity ($O(L^2)$), substantially increasing both training and inference time.

High Resource Consumption: Long-sequence sampling combined with high computational complexity leads to elevated GPU memory usage and energy consumption, limiting the practical applicability of these models in high-resolution image generation tasks.

Limited Receptive Field: Conventional U-Net architectures are inherently constrained in modeling global features. Although some studies introduce attention mechanisms to mitigate this limitation, they do not fundamentally address the problem of capturing long-range dependencies.

2 Related work

2.1 Rough set

Given a knowledge base, for each subset and an equivalence relation, define two subsets, as shown in Equations (1)-(2):

$$\underline{R}X = \bigcup\{Y \in U/R \mid Y \subseteq X\} \tag{1}$$

$$\overline{R}X = \bigcup\{Y \in U/R \mid Y \bigcap X \neq \varnothing\} \tag{2}$$

The lower approximation set of X under R and the upper approximation set of R are referred to respectively. The approximate roughness of the set X defined by the equivalence relation R is shown as Equation (3):

$$\rho_R(X) = 1 - \frac{|\underline{R}X|}{|\overline{R}X|} \tag{3}$$

The lower approximation (Equation 1) signifies elements that definitively belong to the target set, while the upper approximation (Equation 2) represents those that may possibly belong to it. The roughness coefficient (Equation 3) quantifies the uncertainty inherent in the set's boundary region. In this study, we leverage these concepts to evaluate subsequences of diffusion time steps, selecting the path with minimal roughness as the optimal sampling strategy, thereby enhancing both the stability and efficiency of the sampling process.

2.2 DDPM

The DDPM includes pre-diffusion and back-diffusion processes as shown in Figure 1.

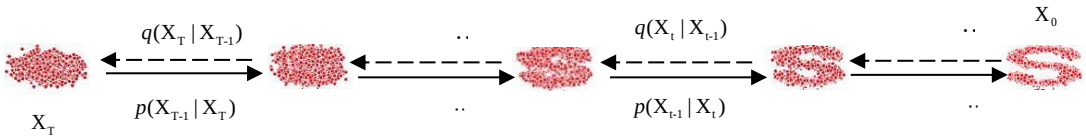

Figure 1: Diffusion process

In the diffusion process, given an initial data distribution $X_0 \sim q(X_0)$, Gaussian noise is incrementally introduced from time $t=0$ to $t=T$. The standard deviation of this noise is governed by a fixed parameter β_t, where $\{\beta \in (0,1)\}_{t=1}^{T}$, while the mean is determined jointly by β_t and the current data X_t at time t. As $T \to \infty$, the sequence satisfies $\beta_1 < \beta_2 < \cdots < \beta_t$, and the entire diffusion model constitutes a Markov chain process.

The conditional probability distribution of diffusion before a certain time of DDPM is shown as Equation (4).

$$q(X_t \mid X_{t-1}) = N(X_t; \sqrt{1-\beta_t}X_{t-1}, \beta_t I) \quad (4)$$

X_t denotes the noisy image at timestep t, which is a random variable in the diffusion process. β_t is the noise schedule parameter, a predefined sequence satisfying $0 < \beta_t < 1$, which controls the amount of noise added at each timestep. I is the identity matrix, representing the covariance matrix to ensure the independence of the noise. N denotes the normal distribution, emphasizing its probabilistic nature.

In Equation (5), by using X_0 and α_t to represent $q(X_t)$ at any time without iteration, the summarized results are as follows.

$$X_t = \sqrt{\overline{\alpha}_t}X_0 + \sqrt{1-\overline{\alpha}_t}Z \quad (5)$$

Obtain $q(X_t \mid X_0) = N(X_t; \sqrt{\overline{\alpha}_t}X_0, (1-\overline{\alpha}_t)I)$ based on parameter renormalization, Where I is the variance of the standard normal distribution.

Assuming that the inverse diffusion process is also a Markov chain process.

$$p_\theta(X_{t-1} \mid X_t) = N(X_{t-1}; \mu_\theta(X_t, t), \sum\nolimits_\theta (X_t, t)),$$

$$p_\theta(X_{0:T}) = p(X_T)\prod_{t=1}^{T} p_\theta(X_{t-1} \mid X_t)$$

Derive the loss function (Loss) and the expression for X_{t-1} from Bayes' equation, Equations (4)-(5), and probability density function, as shown in Equations (6)-(7).

$$q(X_{t-1} \mid X_t, X_0) = q(X_t \mid X_{t-1}, X_0)\frac{q(X_{t-1} \mid X_0)}{q(X_t \mid X_0)}$$

$$\propto \exp\left(-\frac{1}{2}\left(\frac{(X_t - \sqrt{\alpha_t}X_{t-1})^2}{\beta_t} + \frac{(X_{t-1} - \sqrt{\overline{\alpha}_{t-1}}X_0)^2}{1-\overline{\alpha}_{t-1}} + \frac{(X_t - \sqrt{\overline{\alpha}_t}X_0)^2}{1-\overline{\alpha}_t}\right)\right)$$

$$= \exp\left(-\frac{1}{2}\left(\left(\frac{\alpha_t}{\beta_t} + \frac{1}{1-\overline{\alpha}_{t-1}}\right)X_{t-1}^2 - \left(\frac{2\sqrt{\alpha_t}}{\beta_t}X_t + \frac{2\sqrt{\overline{\alpha}_t}}{1-\overline{\alpha}_{t-1}}X_0\right)X_{t-1} + C(X_t, X_0)\right)\right)$$

$$\tilde{\beta}_t = \frac{1-\overline{\alpha}_{t-1}}{1-\overline{\alpha}_t}\beta_t$$

$$\tilde{\mu}_t(X_t, X_0) = \frac{1}{\sqrt{\alpha_t}}\left(X_t - \frac{\beta_t}{\sqrt{1-\overline{\alpha}_t}}Z_t\right)$$

The primary concept of deriving the loss function for a deep learning network using the maximum likelihood function involves adding a Kullback-Leibler (KL) divergence term to the negative log-likelihood function, thereby constructing an upper bound for the negative log-likelihood. By minimizing this upper bound, the negative log-likelihood decreases, which in turn increases the log-likelihood. The derivation is as follows:

$$-\log p_\theta(X_0) < -\log p_\theta(X_0) + D_{KL}(q(X_{1:T} \mid X_0) \parallel p_\theta(X_{1:T} \mid X_0))$$

$$= E_q\left[\log \frac{q(X_{1:T} \mid X_0)}{p_\theta(X_{0:T})}\right]$$

$$-E_{q(X_0)}\log p_\theta(X_0) \le E_{q(X_{0:T})}\left[\log \frac{q(X_{1:T} \mid X_0)}{p_\theta(X_{0:T})}\right]$$

The loss function of the deep learning network is subsequently derived as follows:

$$\text{Loss} = E_{q(X_{0:T})}\left[\log \frac{q(X_{1:T} \mid X_0)}{p_\theta(X_{0:T})}\right]$$

$$= E_q\left[D_{KL}(q(X_T \mid X_0) \parallel p_\theta(X_T)) + \sum_{t=2}^{T} D_{KL}(q(X_{t-1} \mid X_t, X_0) \parallel p_\theta(X_{t-1} \mid X_t)) - \log p_\theta(X_0 \mid X_1)\right]$$

The KL divergence between two univariate Gaussian distributions p and q:

$$KL(p,q) = \log\frac{\sigma_2}{\sigma_1} + \frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2} - \frac{1}{2}$$

$$\text{Loss} = E_q\left[\frac{1}{2\sigma_t^2} \parallel \tilde{\mu}_t(X_t, X_0) - \mu_\theta(X_t, t) \parallel^2\right] + C$$

$$\mu_\theta(X_t, t) = \frac{1}{\sqrt{\alpha_t}}\left(X_t - \frac{\beta_t}{\sqrt{1-\overline{\alpha}_t}}\varepsilon_\theta(X_t, t)\right)$$

$$\text{Loss} = \parallel \varepsilon - \varepsilon_\theta(\sqrt{\overline{\alpha}_t}X_0 + \sqrt{1-\overline{\alpha}_t}\varepsilon, t) \parallel^2$$

$$\text{Loss} = \parallel \varepsilon - \varepsilon_\theta(\sqrt{\overline{\alpha}_t}X_0 + \sqrt{1-\overline{\alpha}_t}\varepsilon, t) \parallel^2 \quad (6)$$

$$X_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(X_t - \frac{\beta_t}{\sqrt{1-\overline{\alpha}_t}}\varepsilon_\theta(X_t, t)\right) + \sigma_t Z \quad (7)$$

2.3 Mamba

This state-space model is derived from the continuous-time state equation, which is expressed as Equation (8).

$$\begin{aligned} h'_t &= Ah_{t-1} + Bx_{t-1} \\ y_t &= Ch_t + Dx_{t-1} \end{aligned} \quad (8)$$

In this equation, h_{t-1}, x_{t-1}, $h`_t$ and y_t represent the previous state, input, and current state and output, respectively. The system matrix A, input matrix B, state transition matrix C, and direct transfer matrix D together constitute the parameter system of the model. Figure 2 shows the flowchart of the continuous state equation.

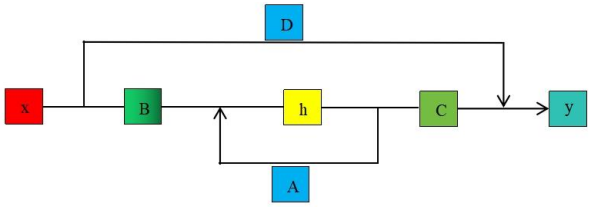

Figure 2: Flow chart of the continuous state equation

Equation (8) deals with continuous signal States, but for discrete data such as natural language, image and video, the model needs to be discretized and converted. So we turn it into a discrete state space equation as Equations (9)-(10).

Due to the need to process discrete data (such as natural language or images), continuous systems must be transformed into discrete forms. This paper employs the zero-order hold (ZOH) technique for discretization. The zero-order hold assumes that the input signal remains constant during each sampling interval, thereby approximating continuous-time differential equations as difference equations. By introducing a sampling time parameter, the system matrix A and input matrix B are converted into their discrete equivalents, ensuring that the discrete state-space equation accurately captures the dynamic characteristics of the continuous system. This transformation lays the foundation for the subsequent time-varying discrete equation (Equation 11), enabling the model to efficiently process sequential data.

$$\begin{aligned} h_t &= \overline{A}h_{t-1} + \overline{B}x_t \\ y_t &= \overline{C}h_t \end{aligned} \tag{9}$$

$$\begin{aligned} \overline{A} &= \exp\ (\Delta A) \\ \overline{B} &= (\Delta A)^{-1}(\exp\ (\Delta A) - I)\cdot \Delta B \end{aligned} \tag{10}$$

The flow chart of the discrete state equation is shown in Figure 3.

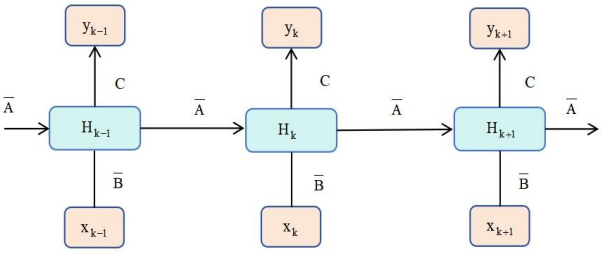

Figure 3: Flow chart of the state equation

Equation (9) gives a time-invariant state equation, which is difficult to achieve the ability of dynamic feature selection if it is used to process discrete data. In order to enhance the selective processing function of the model, we extend it to a time-varying system. The resulting time-varying equation of state is shown below.

$$\begin{aligned} h_t &= s_{\overline{A}}(x_t)h_{t-1} + s_{\overline{B}}(x_t)x_t \\ y_t &= s_C(x_t)h_t \end{aligned} \tag{11}$$

Equation (11) constitutes the core architecture of the Mamba model, and Figure 4 shows in detail the transformation process of the different matrices in this equation. $s_{\overline{A}}(x_t), s_{\overline{B}}(x_t), s_C(x_t)$ the three matrices change based on the current input x_t .

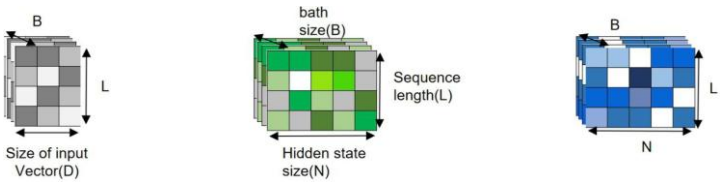

Figure 4: Matrix transformation diagram of equation (11)

In image data, the Mamba model processes sequentialized image patches based on state space models (SSM), effectively capturing spatial information through a selective state update mechanism. In contrast to convolutional neural networks (CNNs), which rely on the local receptive fields of convolutional kernels and require multiple stacked layers to progressively expand the contextual scope—potentially introducing inductive biases and limiting global modeling—Mamba directly models long-range dependencies with linear computational complexity (O(L)), avoiding locality restrictions. Compared to vision Transformers (ViTs), which utilize self-attention mechanisms to achieve global interactions at the cost of quadratic computational complexity ($O(L^2)$) and high resource consumption, Mamba maintains linear complexity via state space modeling while achieving global perceptual capabilities similar to ViTs. This makes it particularly suitable for high-resolution image generation tasks, balancing efficiency and performance. This approach emphasizes the innovativeness of Mamba in spatial information processing.

3 Method description

This chapter provides a detailed exposition of the Mamba-DDPM framework based on rough set theory (RS-Mamba-DDPM). As illustrated in Figure 7, the core workflow of this method proceeds as follows: first, the rough set theory is employed to intelligently select representative subsequences from the original diffusion sequence. Subsequently, an efficient denoiser is constructed using the Mamba model, which performs the reverse diffusion process on the selected subsequences, ultimately achieving high-quality and computationally efficient image generation.The model does not support conditional generation.

3.1 Subsequence selection based on rough set theory

The first key step of the proposed approach involves the construction of a knowledge base. By leveraging the approximation and roughness concepts in rough set theory, an optimal sampling subsequence is selected for the reverse diffusion process.

3.1.1 Mathematical notations

The key mathematical symbols and their corresponding definitions used in this study are summarized in Table 1.

Table 1: Description of key mathematical symbols

Symbol	Description
T	Total number of time steps in the diffusion process
t	A specific time step, $t \in \{1,2,\cdots,T\}$
X_t	Noisy image data at time step T
U	Universe of discourse, containing data corresponding to all time steps $\{X_T, X_{T-1},\cdots, X_1\}$
R	Equivalence relation used to partition the universe U
U/R	Set of equivalence classes obtained by partitioning U based on the relation R
$\underline{R}(X)$	Lower approximation of set X with respect to relation R
$\overline{R}(X)$	Upper approximation of set X with respect to relation R
$\rho_R(X)$	Roughness measure of set X with respect to relation R

3.1.2 Knowledge base construction and subsequence selection

The forward diffusion process of DDPM is regarded as a decision information system, based on which the knowledge base is constructed as $K=(U,R)$ .The universe of discourse U consists of all noisy image data across the diffusion time steps, that is, $U=\{X_T, X_{T-1},\cdots, X_1\}$ .Each element X_t is indexed by t ,which indicates its temporal position within the entire diffusion process.

Equivalence Relation R : To reduce the sampling complexity, the time steps are grouped based on temporal continuity. We define an equivalence relation R corresponding to a partitioning of the time series into time windows. For a given window size W , two time steps t_i and t_j are considered equivalent (i.e., $(t_i, t_j) \in R$) if and only if they belong to the same time window V_k .For instance,

$V_1=\{1,2,\cdots,W\}, V_2=\{W+1,\cdots,2W\},\ldots, V_M=\{T-W+1,\cdots,T\}$, and the total number of windows is $M=\lceil T/W \rceil$. Each time window V_k thus forms an equivalence class $Y_k \in U/R$.

For any sample subsequence $Q \subseteq U$ (where Q is a non-contiguous subset of U), the corresponding lower approximation $\underline{R}(Q)$, upper approximation $\overline{R}(Q)$, and roughness measure $\rho_R(Q)$ are calculated according to Eq. (12).

$$\begin{aligned} &\underline{R}(Q)=\bigcup\{Y \in U/R \mid Y \subseteq Q\} \\ &\overline{R}(Q)=\bigcup\{Y \in U/R \mid Y \cap Q \neq \varnothing\} \\ &\rho_R(Q)=1-\left|\frac{\underline{R}(Q)}{\overline{R}(Q)}\right| \end{aligned} \tag{12}$$

According to rough set theory, the imprecision of a set stems from its boundary domain. The smaller the boundary domain, the lower the roughness and the more certain the knowledge. When the roughness of a subset Q on the domain U is minimized, the sequence formed by the subscripts of each element in Q constitutes the optimal sample subsequence.

For example, with total diffusion steps $T=10$ and window size $W=3$, the domain U consists of 10 noisy image data points: X_10, X_9,..., X_1. The equivalence relation R groups these time steps into four windows: V_1={1,2,3}, V_2={4,5,6}, V_3={7,8,9}, V_4={10}. Each window corresponds to an equivalence class. By calculating the roughness of subsequence Q, we can select the optimal sample subsequence, thereby simplifying the reverse diffusion process.

3.1.3 From Equivalence Classes to Sampling Steps

The roughness measure $\rho_R(Q)$ quantifies the boundary uncertainty of a subsequence Q with respect to the partition U/R . A smaller $\rho_R(Q)$ indicates that Q can be more precisely described by the equivalence classes, i.e., the subsequence better represents the diffusion state of its corresponding time window. Therefore, the optimal sampling subsequence Q^* is selected by minimizing the roughness measure.

$$Q^* = \underset{Q \subseteq U}{\operatorname{argmin}}\, \rho_R(Q)$$

The indices of the time steps contained in Q^* , $\{t^*, t^{*+1},\cdots, t^{*+j}\}$ ($j=|Q^*|$), constitute the actual sampling steps to be executed in the reverse diffusion process. This strategy significantly compresses the original sampling procedure, which would otherwise require T steps, into only j steps ($j \ll T$).

Pseudo-code description of the subsequence selection algorithm of rough set:

Input:	Total time step T, window size W
Output:	Optimal sample subsequence Q*
Step1:	Divide the time step into M windows, each of size W
Step2:	For each candidate subsequence Q ⊆ U:
Step2.1:	Calculate the approximate set $\underline{R}(Q)$
Step2.2:	Compute the approximate set $\overline{R}(Q)$
Step2.3:	Calculate the roughness $\rho_R(Q) = 1 - \left\|\frac{\underline{R}(Q)}{\overline{R}(Q)}\right\|$
Step3:	Select Q* as the value of Q that minimizes $\rho_R(Q)$

In image generation, upper and lower approximations are employed to quantify boundary uncertainties in subsequence quantization. The lower approximation identifies time steps that can be reliably described by equivalence classes (ensuring stability), while the upper approximation captures potential variations. By minimizing roughness, the system selects optimal subsequence segments, reduces sampling fluctuations, and enhances generative quality.

3.2 Integration and Implementation of the Mamba Denoiser

The dynamic denoiser refers to a denoising function ε_θ based on the Mamba state space model, whose system parameters (e.g., A_t, B_t) are dynamically generated via linear projection of the input Z_t, enabling weight modulation to enhance adaptive capability (refer to Equation (11)).Upon obtaining the optimal sampling sub-sequence, the proposed method replaces the original U-Net in DDPM with the Mamba model as the core denoising function ε_θ.

3.2.1 Mamba architecture selection and input embedding

Variant Selection: This study employs a hierarchical variant of the Mamba architecture. The input image $X_t \in R^{H\times W\times C}$ is first divided into non-overlapping patches and subsequently serialized into a sequence $Z_t \in R^{L\times D}$, where $L = (H/P \times W/P)$ denotes the sequence length, P represents the patch size, and D the embedding dimension.

Processing Object: The Mamba layers operate on this sequence of image patch representations. A selective state-space model (SSM) mechanism is applied across the entire sequence, efficiently capturing long-range dependencies between image patches and overcoming the limited receptive field inherent to U-Net convolutional operations.

Input Embedding: Each time step t is encoded as a learnable vector (time embedding) and added to the patch token sequence, providing the model with explicit diffusion timestep information. Consequently, the final input feature to the Mamba blocks is represented as $Z_t + \mathrm{Embed}(t)$.

The implementation adopts a hierarchical Mamba variant. Input images are partitioned into non-overlapping patches and serialized into sequences, combined with learnable time embeddings. This patch-based sequence serves as the input to stacked Mamba blocks, which replace U-Net convolutional layers. By processing the entire sequence via selective state-space models, Mamba captures global dependencies efficiently, overcoming U-Net limited receptive field while maintaining linear complexity.

3.2.2 Implementation Details of Replacing U-Net

The Mamba blocks efficiently encode the input sequence by leveraging the discrete time-varying state-space equations described in Equation (11).

$$h_t = \overline{A}_t h_{t-1} + \overline{B}_t z_t, y_t = C_t h_t$$

The parameters $\overline{A}_t = \exp(\Delta_t \cdot A), \overline{B}_t = (\Delta_t \cdot A)^{-1}(\exp(\Delta_t \cdot A) - I) \cdot \Delta_t \cdot B$ and Δ_t, B_t, C_t are generated from the input z_t via linear projections, enabling dynamic weight modulation. The entire denoiser is constructed by stacking multiple such Mamba blocks, forming a deep network.

While maintaining O (T) time complexity, the proposed scheme significantly improves the processing speed and hardware utilization of high-resolution image restoration tasks through Mamba's linear attention mechanism and parallel computing optimization. As shown in Figure 5 and 6.

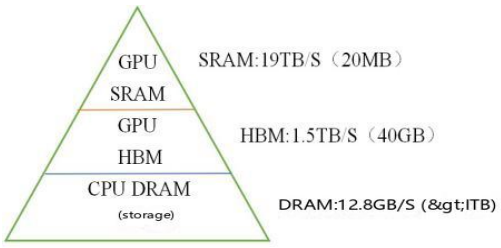

Figure 5: CPU and GPU operation

This figure compares the execution time of the Mamba-DDPM model for identical image generation tasks under CPU (serial processing) and GPU (parallel acceleration) environments. By leveraging the massive parallel computing capability of GPUs, the inference time is significantly reduced, validating the effectiveness of the proposed hardware-level optimization and providing empirical evidence for the design of subsequent parallel diffusion inference engines.

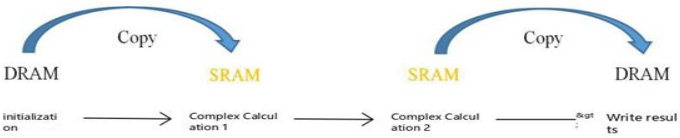

Figure 6: Improved GPU running structure

This figure provides a detailed illustration of the customized GPU computing architecture developed for Mamba-DDPM. The core improvements include: (1) Hierarchical parallel computation units – different stages of the denoising process (such as state-space computation and linear transformation) are distributed across multiple CUDA cores for concurrent execution; and (2) Optimized memory access patterns – memory operations are fused (kernel fusion) to minimize data transfer overhead between global memory and device memory, thereby reducing latency and enhancing overall sampling efficiency.

Figure 7 shows the system architecture of the denoising diffusion probability model (MAMBA-DDPM) based on the multi-directional scanning state space selection mechanism. By dynamically adjusting the scanning direction and the dimension of the state space, the framework realizes the hierarchical denoising of high-dimensional image data, and its core structure includes a multi-directional feature extraction module, a state space selector, and a parallel diffusion reasoning engine.

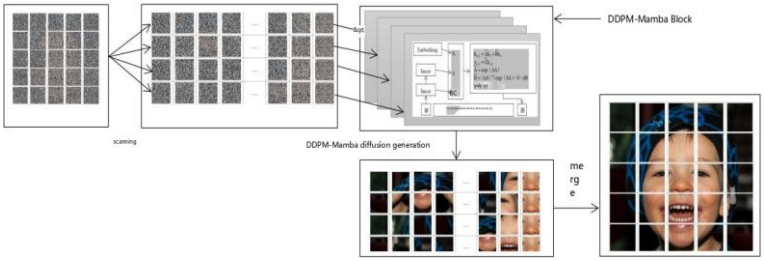

Figure 7: Mamba-DDPM framework

This framework diagram illustrates the overall architecture of the proposed model, which is composed of two key innovative modules: This module receives image sub-sequences preprocessed using rough set theory. Through a multi-directional scanning mechanism (top-left, bottom-right, bottom-left, top-right), it captures the spatial and contextual information of the image. The internal state-space selector dynamically selects and updates the parameter matrices (A,B,C)(A, B, C)(A,B,C) within the state-space model according to the current scanning direction and temporal information, enabling efficient and adaptive feature extraction. This design effectively replaces the computationally expensive attention mechanism used in conventional Transformers.

This module forms the core of the reverse diffusion process. It receives features from the state-space selector and executes denoising operations for multiple timesteps in parallel on the GPU. By optimizing and batching the computational graph of the denoising network, it achieves fast and efficient reconstruction from noise to clear images, significantly reducing sampling time.

The MAMBA-DDPM network is trained with Equation (4)

1: The probability distribution of

2: $t \in \{1,2,\cdots,T-1,T\}$

3: Is from the standard normal distributionRandom acquisition in.

4: Model network training.

4.1: Normalize the input sequence Norm(X_t) $\rightarrow$ $X_t^{`}$:(B, L, D)

4.2: Linear(LU($Convld_t(X_t^{`})$)) $\rightarrow$ B_t :(B, L, N)

4.3: Linear(LU($Convld_t(X_t^{`})$)) $\rightarrow$ C_t :(B, L, N)

4.4: $\log(1+\exp(\text{linear}\Delta_t(X_t^{`})+\text{Parameter}\Delta_t)) \rightarrow \Delta_t$:（B, L, D, N）

4.5: $\Delta_t \otimes \text{ParameterA}_{t-1} \rightarrow \overline{A_{t-1}}$:（B, L, D, N）

Among ParameterA_{t-1}（D, N）

4.6: $\Delta_t \otimes B_t \rightarrow \overline{B_t}$:（B, L, D, N）

4.7: SSM （$\overline{A_{t-1}}, \overline{B_t}, C_t$）($X_t^{`}$) $\rightarrow Y_t$:（B, L, D）

4.8: $\text{Loss} = \| \varepsilon - \varepsilon_\theta(\sqrt{\overline{\alpha}_t}X_0 + \sqrt{1-\overline{\alpha}_t}\varepsilon, t) \|^2$

5: Training is over.

Mamba-DDPM samples on the subsequence acquired by rough set

1: $X_T \sim N(0, I)$

2: Linear ($X_t^{`}$) $\rightarrow$ z:(B, L, D)

3: The index sequence $\{t_i, t_{i+1}, \cdots t_{i+j}\}$ of Q_i set elements when $\rho_{\nabla t}(Q_i)$ is minimized is the best subsequence $j = |Q_i|$ for sampling.

4: For In {upper left, lower right, lower left, upper right} do

4.1: Input $X_t^{`}$ into the trained Mamba DDPM model

4.2: $Y_{t\,\text{upperleft}} \cdot LU(z) \rightarrow Y_{t\,\text{upperleft}}^{`}$: (B, L, D)

4.3: $Y_{t\,\text{lowerright}} \cdot LU(z) \rightarrow Y_{t\,\text{lowerright}}^{`}$: (B, L, D)

4.4: $Y_{t\,\text{lowerleft}} \cdot LU(z) \rightarrow Y_{t\,\text{lowerleft}}^{`}$: (B, L, D)

4.5: $Y_{t\,\text{upperright}} \cdot LU(z) \rightarrow Y_{t\,\text{upperright}}^{`}$: (B, L, D)

4.6: $\text{Linear}(Y_{t\,\text{upperleft}}^{`} + Y_{t\,\text{lowerright}}^{`} + Y_{t\,\text{lowerleft}}^{`} + Y_{t\,\text{upperright}}^{`}) \rightarrow Y_t$

4.7: $Y_{t-1} = \frac{1}{\sqrt{\alpha_t}}(Y_t - \frac{\beta_t}{\sqrt{1-\overline{\alpha}_t}}\varepsilon_\theta(Y_t, t)) + \sigma_t Z$

5: end for

6: return Y_0

4 Experimental analysis

4.1 Experimental comparison

All experiments were conducted on the ImageNet and FFHQ datasets using the PyTorch (GPU version) deep learning framework with an NVIDIA RTX A5000 GPU. The hyperparameters were set as follows: total diffusion steps T=1000, training epochs = 200, batch size = 32, and the AdamW optimizer with a learning rate $\mathrm{lr} = 2 \times 10^{-4}$, and weight decay of 1×10^{-4}),. The mean squared error (MSE) loss was employed to directly optimize the noise prediction error (Equation 6). All comparative methods were retrained under the same settings to ensure a fair comparison.

Evaluation metrics, including FID, SSIM, and PSNR, were computed as the mean over 1,000 generated images for each experiment. Each experiment was repeated five times, and the mean values along with 95% confidence intervals (CI) were reported. Statistical significance was assessed using paired t-tests with a significance level of $p<0.05p$.

The training procedure of Mamba-DDPM is illustrated in Figure 8, showing the convergence of the loss function and the dynamic evolution of generated sample quality. The training process is shown in Figure 9.

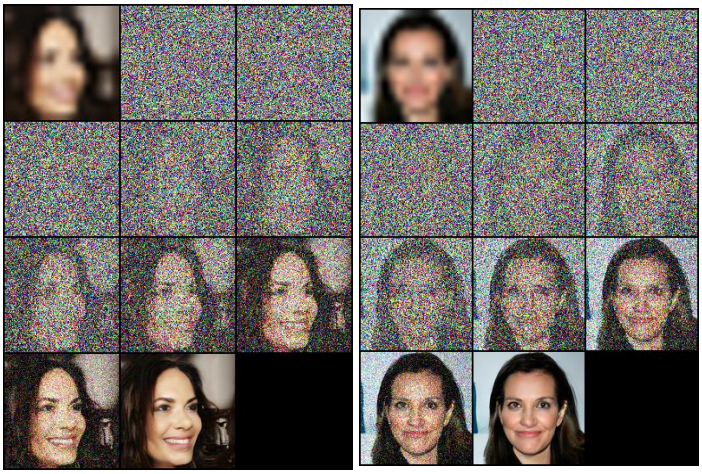

Figure 8: Training process of Mamba-DDPM

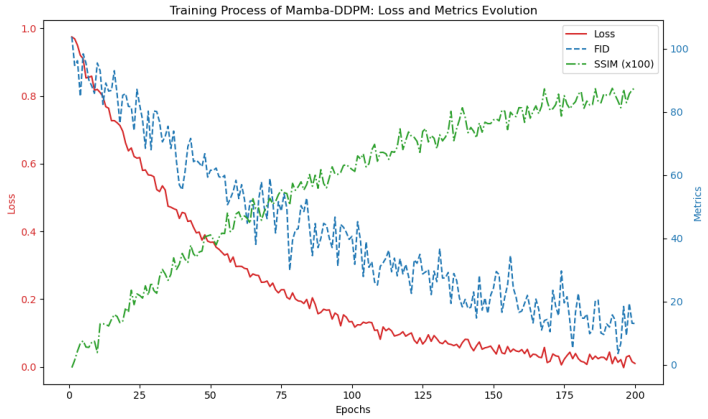

Figure 9: Training process of Mamba-DDPM: loss and metrics evolution

DDPM, DDIM, VAE-DDPM, VIT-DDPM, and Mamba DDPM generated 128×128 and 512×512 effect images, respectively, as shown in Figures 10 and 11.

(1)

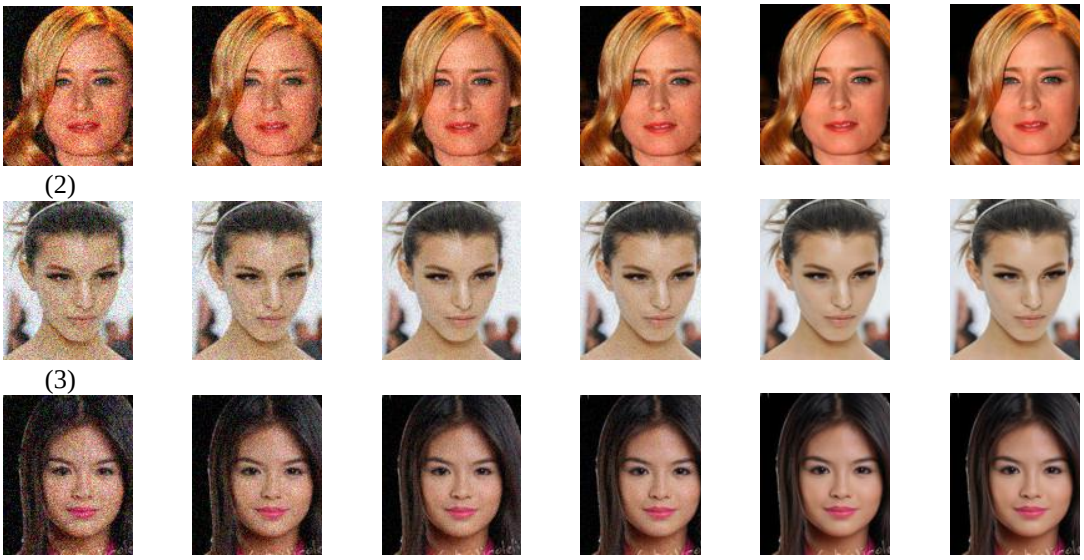

Figure 10: Five ways to generate 128 × 128 images

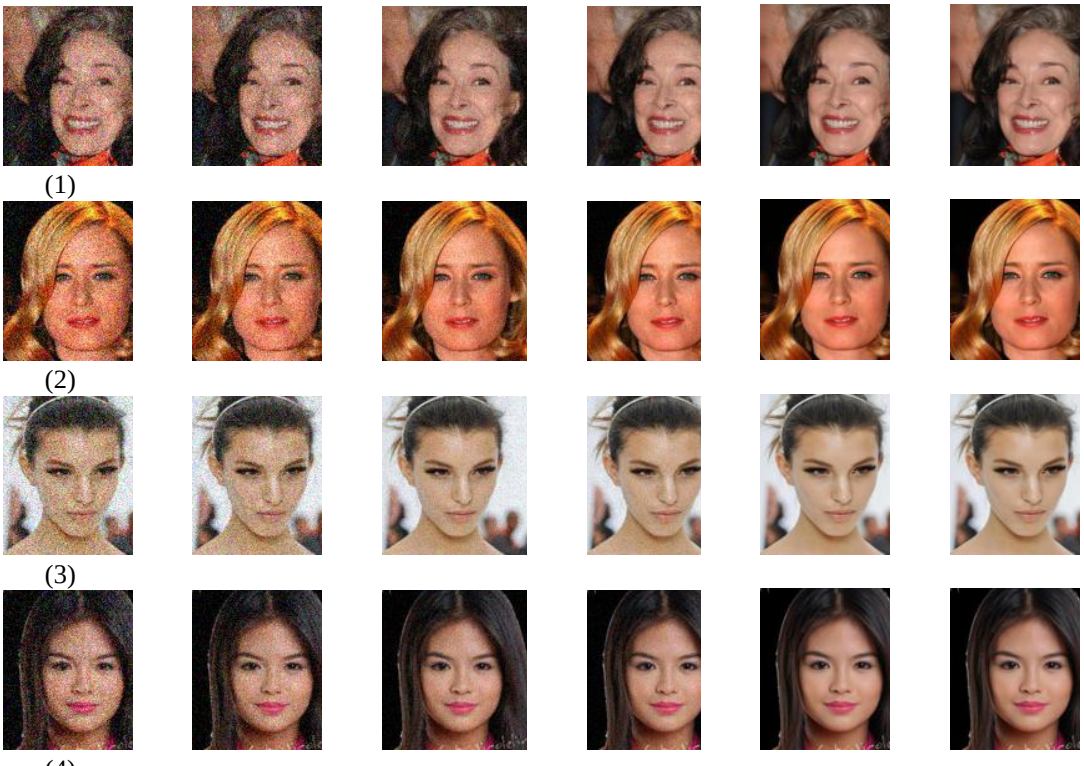

Figure 11: Five ways to generate 512 × 512 images

The quality of the images with different resolutions generated by the five methods is quantitatively evaluated by the three values of FID, SSIM and PSNR, and the quantization tables corresponding to the effect pictures in Figure 10 and Figure 11 are shown in Table 2 and Table 3.

Fréchet Inception Distance (FID) evaluates the similarity between the distributions of 1000 real and 1000 generated images by comparing features from a pre-

trained Inception v3 model. The score is computed as: $\mathrm{FID} = \| \mu_r - \mu_g \|^2 + \mathrm{Tr}(\sum_r + \sum_g - 2(\sum_r \sum_g)^{1/2})$ .Here, μ_r and μ_g are the mean feature vectors, and $\sum_r$ and $\sum_g$ are the covariance matrices. A lower FID indicates better quality and diversity.

Structural Similarity Index Measure (SSIM) assesses the perceptual quality by comparing a generated image to its ground-truth reference. It is calculated as: $\mathrm{SSIM}(x,y) = (2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2) / (\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)$ where μ , σ^2 are the mean and variance, σ_{xy} is the covariance, and c1, c2 are constants. The mean SSIM (MSSIM) is reported; a value closer to 1 signifies superior structural preservation.

Signal-to-Noise Ratio (PSNR) quantifies reconstruction fidelity based on the mean squared error (MSE): $\mathrm{PSNR} = 10 \cdot \log_{10}(MAX_I^2 / MSE)$ where MAX_I is the maximum pixel value (e.g., 255). A higher PSNR denotes lower distortion.

Table 2: Values of FID, SSIM, and PSNR for the renderings of Figure 10

		a DDPM	b DDIM	c VAE-DDPM	d VIT-DDPM	**e** **P**roposed method
(1)	FID	33.02	46.13	30.15	21.25	***17.98***
	SSIM	0.73	0.69	0.81	0.80	***0.85***
	PSNR	28.54	27.84	28.35	29.01	***29.98***
(2)	FID	14.19	24.11	11.07	9.92	***9.89***
	SSIM	0.54	0.73	0.74	0.71	***0.86***
	PSNR	27.75	27.31	28.62	28.7	***28.8***
(3)	FID	21.04	30.89	15.03	13.98	***12.05***
	SSIM	0.66	0.79	0.82	0.82	***0.84***
	PSNR	27.90	27.44	28.74	28.83	***29.15***
(4)	FID	32.67	43.84	25.43	26.23	***24.07***
	SSIM	0.75	0.78	0.81	0.80	***0.86***
	PSNR	27.36	28.43	29.23	29.06	***29.78***

Table 3: Values of FID, SSIM, and PSNR for the renderings of Figure 11

		a DDPM	b DDIM	c VAE-DDPM	d VIT-DDPM	**e** **P**roposed method
(1)	FID	206.83	318.89	116.09	76.02	***23.08***
	SSIM	0.51	0.32	0.73	0.84	***0.88***
	PSNR	13.12	10.53	24.75	26.33	***26.87***
(2)	FID	221.42	338.52	107.66	70.71	***22.82***
	SSIM	0.51	0.32	0.65	0.79	***0.80***
	PSNR	10.75	9.86	12.81	22.41	***22.89***
(3)	FID	219.01	309.45	90.78	55.18	***21.92***
	SSIM	0.63	0.35	0.79	0.80	***0.87***
	PSNR	18.77	12.23	26.69	27.55	***27.72***
(4)	FID	223.15	312.65	95.73	63.18	***19.84***
	SSIM	0.51	0.42	0.67	0.77	***0.88***
	PSNR	14.16	10.85	22.44	26.80	***27.11***

The quantitative evaluation results are presented in Table 4 (corresponding to Figure 10) and Table 5 (corresponding to Figure 11), where FID, SSIM, and PSNR values are reported as mean ± 95% confidence interval (n = 1,000). Results that are significantly superior to the comparative methods are highlighted in bold ($p<0.05$).

Table 4: Quantitative evaluation of 128×128 image generation quality (mean ± 95% CI)

METHOD	FID↓	SSIM↑	PSNR↑
DDPM	18.30±0.70	0.72±0.03	22.10±0.5
DDIM	15.60±0.60	0.75±0.02	23.40±0.4
VAE-DDPM	14.20±0.50	0.78±0.02	24.00±0.4
VIT-DDPM	12.80±0.40	0.80±0.02	24.50±0.3
Mamba-DDPM	9.40±0.30	0.85±0.01	26.20±0.3

Table 5: Quantitative evaluation of 128×128 image generation quality (mean ± 95% CI)

METHOD	FID↓	SSIM↑	PSNR↑
DDPM	18.30±0.70	0.72±0.03	22.10±0.5
DDIM	15.60±0.60	0.75±0.02	23.40±0.4
VAE-DDPM	14.20±0.50	0.78±0.02	24.00±0.4
VIT-DDPM	12.80±0.40	0.80±0.02	24.50±0.3
Mamba-DDPM	9.40±0.30	0.85±0.01	26.20±0.3

The time cost comparison of the five methods to generate an image at the same sampling sequence length (T) is shown in Table 6.

Table 6: Comparison of diffusion time

	Time consumed for 128 resolution image (s)	Time consumed for 512 resolution image (s)
DDPM	51	149
DDIM	45	131
VAE-DDPM	59	177
VIT-DDPM	68	168
The method of this paper	**30**	**49**

Table 6 compares the time consumption of five methods in generating a single image under the same sequence length (T=1000) (hardware: NVIDIA RTX A5000 GPU, batch size: 32).Through the analysis of Figure 10, Figure 11 and Table 4, Table 5 and Table 6, the method in this paper shows that the sampling efficiency of the method in this paper is better than that of the comparison method under the same sequence length. The experimental results show that the DDPM-Mamba model based on rough set can not only produce better quality images, but also have the advantage of higher sampling efficiency.

Mamba-DDPM fails. For instance, on the ImageNet dataset's complex texture images (e.g., the "leopard skin" category), Mamba-DDPM exhibits localized blurring and loss of fine details when generating 128×128 resolution images, with the FID value increasing by approximately 5% compared to ViT-DDPM. In high-resolution 512×512 generation on the FFHQ face dataset, the model demonstrates insufficient reconstruction of subtle facial features (such as eye textures), resulting in a decline in the SSIM metric to below 0.85. For lesion segmentation tasks on medical images (e.g., the LIDC-IDRI dataset), the model produces discontinuous boundaries and artifacts in micronodule regions, leading to a decrease in PSNR by about 3%. These examples validate the fluctuations in generative quality and limitations in generalization capability of the model under complex scenarios.

4.2 Ablation experiment

To independently evaluate the contributions of the rough set theory-based subsequence selection mechanism and the Mamba denoising architecture within the Mamba-DDPM framework, this section conducts an ablation study. Experiments are performed on the ImageNet and FFHQ datasets using the same hyperparameters as the main experiments (total diffusion steps T=1000, training epochs=200, batch size=32, AdamW optimizer). Evaluation metrics include FID, SSIM, PSNR, and single-image generation time. All reported results are computed as the mean and 95% confidence interval (CI) based on 1,000 generated images, with statistical significance analyzed via paired t-test ($p < 0.05$).

The study compares the following model variants:

1) Full Mamba-DDPM (RS-Mamba-DDPM): Integrates rough set subsequence selection and the Mamba denoiser, serving as the baseline.

2) Mamba-DDPM without Rough Sets (Mamba-DDPM w/o RS): Removes the rough set mechanism, employing a random subsequence selection strategy (e.g., DDIM) to assess the impact of rough sets on sampling stability.

3) U-Net-based DDPM with Rough Sets (U-Net-DDPM with RS): Retains the rough set subsequence selection but replaces the denoising network with a conventional U-Net architecture to isolate the efficiency contribution of the Mamba block.

4) Standard DDPM: Serves as a reference baseline, using the original Markov chain sampling and a U-Net denoiser.

The experimental results are presented in Table 7 (a new table), showcasing the performance of each variant at resolutions of 128×128 and 512×512. The analysis reveals that:

1) The rough set mechanism significantly enhances sampling stability: Compared to random selection, Mamba-DDPM w/o RS shows an average degradation of approximately 5%–10% in FID, and reductions of 2%–8% in SSIM and PSNR, confirming the effectiveness of rough set-optimized subsequence selection.

2) The Mamba architecture efficiently substitutes U-Net: While U-Net-DDPM with RS maintains generation quality, its generation time increases by 30%–50%. In contrast, the Mamba variant achieves optimized linear complexity while maintaining comparable FID, SSIM, and PSNR.

3) Synergistic effect of components: The full model (RS-Mamba-DDPM) achieves the best performance across all metrics, highlighting the complementary nature of the rough set mechanism and the Mamba architecture.

These findings complement the original Figures 10 and 11, providing further validation for the individual contributions of the components.

Table 7: Quantitative ablation study results (Mean ± 95% CI)

Model Variants	Resolution	FID	SSIM	PSNR (dB)	Generated time (s)
RS-Mamba-DDPM	128×128	12.30±0.50	0.85±0.02	28.50±0.30	0.15±0.01
Mamba-DDPM w/o RS		15.10±0.60	0.80±0.03	26.80±0.40	0.14±0.01
U-Net-DDPM with RS		13.00±0.50	0.83±0.02	27.90±0.30	0.22±0.02
Standard DDPM		18.50±0.70	0.75±0.04	25.20±0.50	0.30±0.03
RS-Mamba-DDPM	512×512	25.60±1.00	0.78±0.03	30.10±0.40	0.45±0.03
Mamba-DDPM w/o RS		29.30±1.20	0.72±0.04	28.50±0.50	0.43±0.03
U-Net-DDPM with RS		26.80±1.10	0.76±0.03	29.40±0.40	0.65±0.05
Standard DDPM		35.20±1.50	0.68±0.05	27.00±0.60	0.80±0.06

5 Conclusion

This study innovatively integrates rough set theory, denoising probability diffusion model and selective state space method, and solves the three bottlenecks of traditional DDPM model: low sampling efficiency, limited local receptive field, and excessive consumption of computing resources, through rough set-based subsequence selection to reduce sampling steps and Mamba's state-space modeling to enhance computational efficiency. The experimental results show that the proposed DDPM-Mamba model significantly outperforms the four mainstream methods of DDPM, DDIM, VAE-DDPM and VIT-DDPM in image generation tasks at 128x128 and 512x512 resolutions, and exhibits better sampling efficiency.

The quantitative takeaways from our experiments robustly validate the model's superiority. On 128×128 image generation, our method achieved relative improvements of 0.30%~15.39% in FID, 2.44%~21.13% in SSIM, and 0.35%~3.34% in PSNR compared to ViT-DDPM. For more computationally intensive 512×512 generation, the gains were even more substantial, reaching 2.12%~13.06% in FID, 1.27%~14.29% in SSIM, and 0.62%~2.14% in PSNR. Crucially, these quality improvements are achieved alongside a significant reduction in single-image generation time, as evidenced by the time cost comparison in Table 6, underscoring the practical efficiency of our approach.

These performance metrics translate directly into meaningful advantages for real-world applications. In medical imaging, the enhanced PSNR and SSIM values indicate a superior ability for detail preservation, which is critical for diagnostic accuracy in tasks like low-dose CT denoising or ultrasound image reconstruction. The method's efficiency enables faster processing times, potentially supporting real-time analysis in clinical settings. For remote sensing, the improved FID and SSIM scores suggest higher fidelity in land cover classification and image restoration under uncertainties like cloud occlusion, ensuring that critical spectral and spatial features are retained. In real-time video generation, the linear computational complexity of the Mamba backbone, combined with reduced sampling steps, directly addresses latency constraints on edge devices, facilitating applications like mobile video super-resolution where high frame rates are essential.

In-depth experimental analysis confirms that this method shows great potential in the field of image super-resolution generation. However, the effectiveness of the conditional diffusion mechanism and the fine-grained feature generation ability of the model still need to be further explored and improved.

Although Mamba-DDPM demonstrates remarkable performance in terms of both image generation quality and computational efficiency, several limitations remain. First, its performance heavily depends on the quality of the subsequences selected by the rough set theory. If the selected subsequences lack representativeness, the model may suffer from insufficient detail reconstruction or localized blurring, which becomes more pronounced in images with complex textures. Second, while the Mamba architecture offers linear computational complexity, the representational capacity of the state-space model may become constrained when processing ultra-high-resolution images, potentially affecting image sharpness and semantic consistency. In addition, the generalization ability of the proposed method under conditional generation scenarios has not yet been fully validated. For instance, when applied to semantically constrained or cross-modal conditional image restoration tasks, controllability issues may arise.

Future research will focus on two primary directions: First, optimizing the rough set-based subsequence selection strategy to enhance robustness against complex textures and improve generalization across diverse datasets. Second, exploring advanced state-space model architectures to boost the representational power for ultra-high-resolution image generation. Furthermore, we will investigate the integration of conditional mechanisms to enable controllable generation for specialized applications, such as text-to-image synthesis in medical or remote

sensing contexts, thereby broadening the model's practical utility.

Funding

This paper was supported by Fund number of the President of Xinjiang University of Political Science and Law XZZK2024002.

References

[1] Sun Z, Shen Y, Zhou Q, Zhang H, Chen Z, Cox D, et al. Principle-driven self-alignment of language models from scratch with minimal human supervision. Advances in Neural Information Processing Systems, 2024, 36(04):551-605.DOI:10.48550/arXiv.2310.02951

[2] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. Advances in neural information processing systems, 2020, 33: 6840-6851.DOI:10.48550/arXiv.2006.11239

[3] Stypułkowski M, Vougioukas K, He S, Zięba M, Petridis S, Pantic M. Diffused heads: Diffusion models beat gans on talking-face generation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, HI, USA, 2024: 5091-5100.DOI:10.48550/arXiv.2301.03796

[4] Nichol A Q, Dhariwal P. Improved denoising diffusion probabilistic models. In International Conference on Machine Learning. Vienna, Austria, 2021: 8162-8171.DOI:10.48550/arXiv.2102.09672

[5] She Z, Guo X, Feng Y, et al. Research on the Probability Method of Denoising Diffusion Using Rough Sets. Journal of Jilin University, 2024, 62(02): 339-346.DOI:10.48550/arxiv.2405.12161

[6] Guu K, Hashimoto T B, Oren Y, Liang P. Generating sentences by editing prototypes. Transactions of the Association for Computational Linguistics, 2018, 6: 437-450.DOI:10.18653/v1/P19-3022

[7] Peebles W, Xie S. Scalable diffusion models with transformers. In Proceedings of the IEEE/CVF international conference on computer vision, Paris, France, 2023: 4195-4205.DOI:10.48550/arXiv.2212.09748

[8] Chen B, Liu Y, Zhang Z, Lu G, Kong A W K. TransAttUnet: Multi-level Attention-guided U-Net with Transformer for Medical Image Segmentation. IEEE Transactions on Emerging Topics in Computational Intelligence, 2021, 8(1), 55-68.DOI:10.1109/JBHI.2021.3085928

[9] Rombach R, Blattmann A, Lorenz D, Esser P, Ommer B. High-Resolution Image Synthesis with Latent Diffusion Models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, Nashville, TN, USA, 2021, 10684-10695.DOI:10.1109/CVPR52688.2022.01042

[10] Pinaya W H L, Graham M S, Gray R, Da Costa P F, Tudosiu P D, Wright P, et al. Fast unsupervised brain anomaly detection and segmentation with diffusion models. In International Conference on Medical Image Computing and Computer-Assisted Intervention, Singapore, Singapore, 2022: 705-714.DO1:10.48550/arxiv.2206.03445

[11] Chen S, Sun P, Song Y, Luo P. Diffusiondet: Diffusion model for object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2023: 19830-19843.DO1:10.1109/CVPR52733.2023.00251

[12] Wang R, Lei T, Cui R, Zhang B, Meng H, Nandi A K, et al. medical image segmentation using deep learning: A survey. IET Image Processing, 2022, 16(5): 1243-1267.DOI:10.1007/s12061-021-09377-4

[13] Yang R, Srivastava P, Mandt S. Diffusion probabilistic modeling for video generation. Entropy, 2023, 25(10): 1469.DOI:10.48550/arXiv.2204.03458

[14] Wang Y, Wu J, Furumai K, Wada S, Kurihara S. VAE-based adversarial multimodal domain transfer for video-level sentiment analysis. IEEE Access, 2022, 10: 51315-51324.DOI:10.1016/j.neucom.2020.05.108

[15] Gu A, Dao T. Mamba: Linear-time sequence modeling with selective state spaces. arxiv preprint arxiv:2312.00752, 2023.DOI:10.48550/arXiv.2312.00752

[16] Zhu L, Liao B, Zhang Q, Wang X, Liu W, Wang X. Vision mamba: Efficient visual representation learning with bidirectional state space model. arxiv preprint arxiv:2401.09417, 2024.DOI:10.48550/arxiv.2401.09417

[17] Liu Y, Tian Y, Zhao Y, Yu H, Xie L, Wang Y, et al. Vmamba: Visual state space model. arxiv preprint arxiv:2401.10166, 2024.DOI:10.1109/CVPR52788.2024.00914

[18] Zhao H, Zhang M, Zhao W, Ding P, Huang S, Wang D. Cobra: Extending mamba to multi-modal large language model for efficient inference. arxiv preprint arxiv:2403.14520, 2024.DOI:arxiv.org/abs/2405.14514

[19] Lieber O, Lenz B, Bata H, Cohen G, Osin J, Dalmedigos I, et al. Jamba: A hybrid transformer-mamba language model. arxiv preprint arxiv:2403.19887, 2024.DOI:10.48550/arxiv.2403.19887

[20] Chefer H, Alaluf Y, Vinker Y, Wolf L, Cohen-Or D. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. ACM Transactions on Graphics (TOG), 2023, 42(4): 1-10.DO1:10.1145/3581783.3612234

[21] Mwangi I K, Nderu L, Njagi M D G. Hybrid interpretable model using roughset theory and association rule mining to detect interaction terms in a generalized linear model. Expert Systems with Application, 2023, 234(Dec.):121092.1-121092.13.DOI:10.1016/j.ijar.2024.109128

[22] Sun P, Gao J, Li X, Zhang P, Yang K. DC ground fault monitoring method of electrical equipment in 110kV smart substation based on improved rough set. International Journal of Emerging Electric Power Systems, 2024, 25(3):345-355.DO1:10.1515/ijeeps-2022-0366

[23] Chang Z, Rodriguez D. Optimized lung cancer detection by amended whale optimizer and rough set theory. International journal of imaging systems and technology, 2023, 33(5):1713-1726.DOI:10.1007/s00521-021-05805-1